BMC Bioinformatics

BioMed Central

Methodology article

Open Access

ORFer – retrieval of protein sequences and open reading frames from GenBank and storage into relational databases or text files

Konrad Büssow*, Steve Hoffmann and Volker Sievert

Address: Protein Structure Factory, Max Planck Institute of Molecular Genetics, Heubnerweg 6, 14059 Berlin, Germany

Email: Konrad Büssow* - buessow@molgen.mpg.de; Steve Hoffmann - Hoffmann@stud-mailer.uni-marburg.de; Volker Sievert - sievert@molgen.mpg.de

* Corresponding author

Published: 19 December 2002

BMC Bioinformatics 2002, **3**:40

Received: 20 September 2002

Accepted: 19 December 2002

This article is available from: http://www.biomedcentral.com/1471-2105/3/40

Abstract

Background: Functional genomics involves the parallel experimentation with large sets of proteins. This requires management of large sets of open reading frames as a prerequisite of the cloning and recombinant expression of these proteins.

Results: A Java program was developed for retrieval of protein and nucleic acid sequences and annotations from NCBI GenBank, using the XML sequence format. Annotations retrieved by ORFer include sequence name, organism and also the completeness of the sequence. The program has a graphical user interface, although it can be used in a non-interactive mode. For protein sequences, the program also extracts the open reading frame sequence, if available, and checks its correct translation. ORFer accepts user input in the form of single or lists of GenBank GI identifiers or accession numbers. It can be used to extract complete sets of open reading frames and protein sequences from any kind of GenBank sequence entry, including complete genomes or chromosomes. Sequences are either stored with their features in a relational database or can be exported as text files in Fasta or tabulator delimited format. The ORFer program is freely available at http://www.proteinstrukturfabrik.de/orfer.

Conclusion: The ORFer program allows for fast retrieval of DNA sequences, protein sequences and their open reading frames and sequence annotations from GenBank. Furthermore, storage of sequences and features in a relational database is supported. Such a database can supplement a laboratory information system (LIMS) with appropriate sequence information.

Background

The functional characterization of large sets of genes includes the characterization of the encoded proteins. Structural genomics aims at determining structures of large sets of proteins that will represent all domain folds present in the biosphere [1]. The cloning and expression of large sets of open reading frames and proteins requires the management and analysis of significant amounts of data. In the Protein Structure Factory, a collaborative Structural Genomics project http://www.proteinstrukturfabrik.de, a relational database system is used to store sequence information and experimental data on proteins chosen as targets for structure determination. These targets consist of human protein sequence entries of the GenBank protein database. Sequences and annotations of the target proteins are integrated in the database. The ORFer program has been developed to accomplish the input of large sets of protein sequence entries of the GenBank database http://www.ncbi.nlm.nih.gov, together with the corresponding coding DNA sequences, into our relational

database. The open reading frame sequences of the target proteins are subsequently used to automatically design PCR primers and proceed with the expression of the target proteins.

ORFer is able to extract complete sets of open reading frames (ORFs) and protein sequences from any GenBank sequence entry, including complete genomes or chromosomes.

The tabular data that ORFer generates is easily accessible to statistical calculation, e.g. to determine the distribution of sequence length or the proportion of incomplete sequence entries in larger sets of sequences.

Completeness of GenBank database entries

ORFer extracts information on the "completeness" of ORFs from GenBank. The term "completeness" in GenBank sequence entries refers to a biomolecule, a clone, or a biological entity as gene, transcript or open reading frame. A complete ORF ranges from the initiation to the stop codon.

Alternative software

Open reading frames of single protein sequences can also be retrieved directly from the GenBank web site by following the appropriate links in the HTML display of GenBank protein sequences.

Functionality similar to ORFer is offered by the program coderet of the Emboss package http://www.emboss.org. Coderet extracts transcript, open reading frame and protein sequences from GenBank DNA sequence entries. The current version of ORFer extracts ORF and protein sequences, but not transcript sequences. Coderet does not interpret GenBank protein sequence entries. Installation of the Emboss package on a local UNIX server is required to run coderet; ORFer is a stand-alone application, but can easily be integrated into LIMS applications using its database interface. The Emboss package was originally designed for command-line use, but graphical interfaces are being developed, e.g. the Jemboss interface. Coderet is fully integrated with the Emboss package. ORFer can save sequences in Fasta format, the default sequence format of the Emboss package.

GenBank XML sequence format

Sequences can be retrieved from GenBank in a variety of formats, including the XML http://www.w3.org/XML/ and ASN.1 formats. Thereby sequence annotations are stored in an structured, tree-like fashion which allows for efficient information retrieval using a parsing process. Both formats share the same data tree structure, and a description of the ASN.1 format also applies to the structure of the XML files [2], http://www.ncbi.nlm.nih.gov/IEB/ToolBox/XML/ncbixml.txt.

Single GenBank XML files often contain sets of sequences. In the case of DNA sequences containing ORFs that encode protein sequences, these DNA and protein sequences are usually included in the same XML file. Sets of DNA sequences, e.g. the exon sequences of a gene, are also combined in single XML files.

Results and Discussion

ORFer was developed to retrieve DNA sequences, protein sequences and their open reading frames and sequence annotations from GenBank and to store them in a relational database. We choose to use the XML sequence format of GenBank, since public domain XML parsers are available for Java, and open reading frames are stored in standard manner in the GenBank XML files. The XML files retrieved for protein sequences usually contain the encoding DNA sequence.

Only GenBank sequence can be read by the current version of ORFer. Since GenBank includes protein sequences of PDB and SwissProt, and is synchronised to the EMBL database, this might not impose a serious restriction.

The Ensembl genome database [3] offers a publicly available mySQL database server, and therefore XML or flat file parsing software is not required to extract information from Ensembl to a local database. Instead, protein and open reading frame sequences and annotations can be directly obtained by suitable SQL queries.

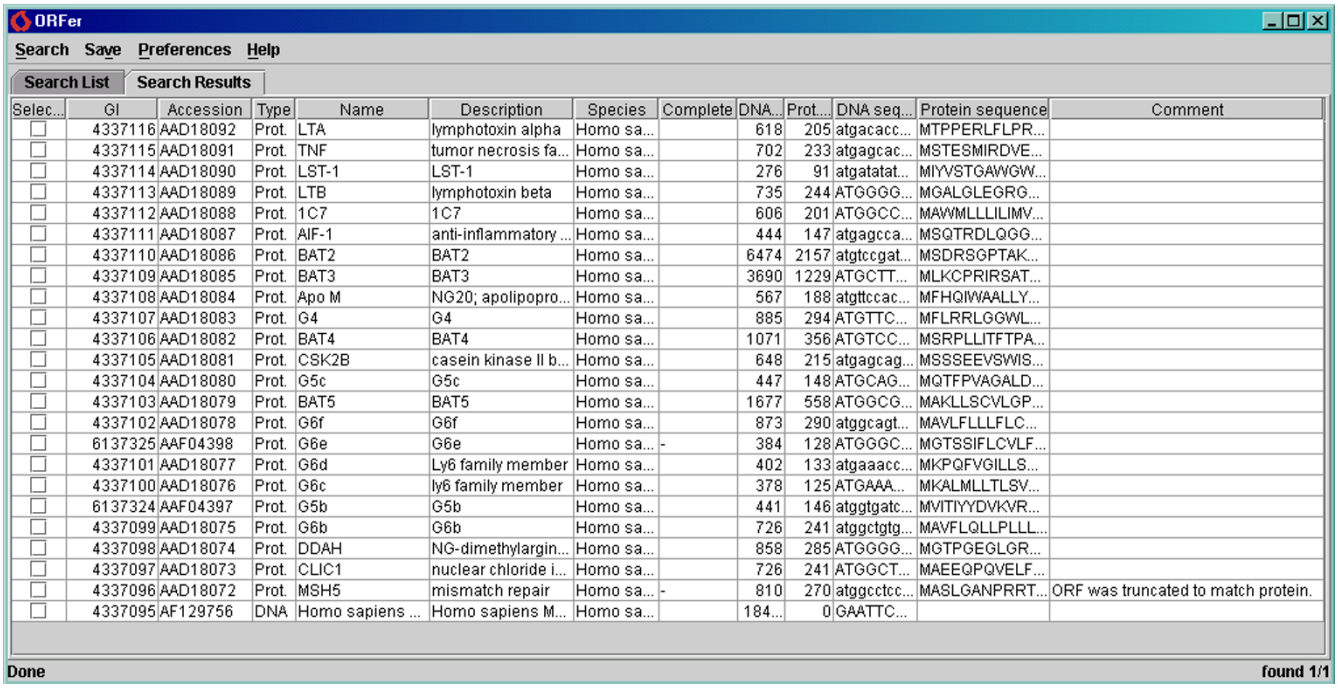
ORFer retrieves sequence entries by either their GenBank molecular biology database identifier (GI) or accession numbers. The user supplies these identifiers, either interactively or as text files containing lists of identifiers.

Sequences are stored with their features in a relational database (mySQL, Oracle or Microsoft Access) or can be exported as text files in Fasta or tab-delimited format.

ORFer can be used to visually inspect hundreds of protein or DNA sequence entries in a tabular view. Another option is to display all sequences found in a single XML file, e.g. all open reading frames contained in a genomic DNA sequence entry (Figure 1).

Database schema

A uniform database schema was implemented for the DBMS mySQL, Oracle8 or greater and Microsoft Access 97. The database has a main table GENBANKENTRIES, and three associated tables for species information and sequence data (Figure 2). The GENBANKENTRIES table



Selec...	GI	Accession	Type	Name	Description	Species	Complete	DNA...	Prot...	DNA seq...	Protein sequence	Comment
☐	4337116	AAD18092	Prot.	LTA	lymphotoxin alpha	Homo sa...		618	205	atgacacc...	MTPPERLFLPR...	
☐	4337115	AAD18091	Prot.	TNF	tumor necrosis fa...	Homo sa...		702	233	atgagcac...	MSTESMIRDVE...	
☐	4337114	AAD18090	Prot.	LST-1	LST-1	Homo sa...		276	91	atgatatat...	MIYVSTGAWGW...	
☐	4337113	AAD18089	Prot.	LTB	lymphotoxin beta	Homo sa...		735	244	ATGGGG...	MGALGLEGRG...	
☐	4337112	AAD18088	Prot.	1C7	1C7	Homo sa...		606	201	ATGGCC...	MAWMLLLILIMV...	
☐	4337111	AAD18087	Prot.	AIF-1	anti-inflammatory ...	Homo sa...		444	147	atgagcca...	MSQTRDLQGG...	
☐	4337110	AAD18086	Prot.	BAT2	BAT2	Homo sa...		6474	2157	atgtccgat...	MSDRSGPTAK...	
☐	4337109	AAD18085	Prot.	BAT3	BAT3	Homo sa...		3690	1229	ATGCTT...	MLKCPRIRSAT...	
☐	4337108	AAD18084	Prot.	Apo M	NG20; apolipopro...	Homo sa...		567	188	atgttccac...	MFHQIWAALLY...	
☐	4337107	AAD18083	Prot.	G4	G4	Homo sa...		885	294	ATGTTC...	MFLRRLGGWL...	
☐	4337106	AAD18082	Prot.	BAT4	BAT4	Homo sa...		1071	356	ATGTCC...	MSRPLLITFTPA...	
☐	4337105	AAD18081	Prot.	CSK2B	casein kinase II b...	Homo sa...		648	215	atgagcag...	MSSSEEVSWIS...	
☐	4337104	AAD18080	Prot.	G5c	G5c	Homo sa...		447	148	ATGCAG...	MQTFPVAGALD...	
☐	4337103	AAD18079	Prot.	BAT5	BAT5	Homo sa...		1677	558	ATGGCG...	MAKLLSCVLGP...	
☐	4337102	AAD18078	Prot.	G6f	G6f	Homo sa...		873	290	atggcagt...	MAVLFLLLFLC...	
☐	6137325	AAF04398	Prot.	G6e	G6e	Homo sa...	-	384	128	ATGGGC...	MGTSSIFLCVLF...	
☐	4337101	AAD18077	Prot.	G6d	Ly6 family member	Homo sa...		402	133	atgaaacc...	MKPQFVGILLS...	
☐	4337100	AAD18076	Prot.	G6c	ly6 family member	Homo sa...		378	125	ATGAAA...	MKALMLLTLSV...	
☐	6137324	AAF04397	Prot.	G5b	G5b	Homo sa...		441	146	atggtgatc...	MVITIYYDVKVR...	
☐	4337099	AAD18075	Prot.	G6b	G6b	Homo sa...		726	241	atggctgtg...	MAVFLQLLPLLL...	
☐	4337098	AAD18074	Prot.	DDAH	NG-dimethylargin...	Homo sa...		858	285	ATGGGG...	MGTPGEGLGR...	
☐	4337097	AAD18073	Prot.	CLIC1	nuclear chloride i...	Homo sa...		726	241	ATGGCT...	MAEEQPQVELF...	
☐	4337096	AAD18072	Prot.	MSH5	mismatch repair	Homo sa...	-	810	270	atggcctcc...	MASLGANPRRT...	ORF was truncated to match protein.
☐	4337095	AF129756	DNA	Homo sapiens ...	Homo sapiens M...	Homo sa...		184...	0	GAATTC...		

Figure 1
Screen shot of the ORFer application. A search for GI identifier 4337095 retrieves all open reading frames and encoded proteins within this sequence entry. Table view of sequences retrieved from GenBank. A set of DNA and protein sequences is shown. For the protein sequences, the DNA sequence column contains the open reading frame sequence.

contains fields for accession number, GI number, sequence name and description, sequence length etc.

XML parsing: protein and gene names and species

Each sequence in a GenBank XML file is represented by an XML element named Bioseq. GI identifier and accession number are found in the Seq-id_gi and Textseq-id_accession elements, respectively. The organism's species and genus are stored in the BinomialOrgName_species and BinomialOrgName_genus elements. For organisms, which do not have a species and genus name, e.g. viruses, the element Org-ref_taxname is used. In the local database, species are stored non-redundantly in the SPECIES table.

ORFer will read the protein or gene name and description from the elements Prot-ref_name_E and Prot-ref_desc, if available, or from a Seqdesc_title element within the respective Bioseq element or finally from a Seqdesc_title element occurring outside the Bioseq elements, if nothing else is available.

XML parsing: sequence completeness

GenBank sequence entries that are known to be complete or incomplete contain MolInfo_completeness elements with an attribute named value. This attribute can have various values, e.g. "no-left" for sequences that are incomplete in the upstream direction, or "complete" for complete sequences. In addition, a seq-feat element representing a translation may contain the element Seq-feat_partial with the attribute value="true", which signifies that a translation product sequence is incomplete.

ORFer looks for both kind of elements to determine whether a sequence entry has been annotated as complete or incomplete. In the "Complete" column of the ORFer Search Result window, a "+" is displayed for complete sequences and a "-" for incomplete sequences. The COMPLETECDS field in the GENBANKENTRIES table of the relational database schema contains either the values 0: "sequence known to be incomplete", or 1: "unknown" or 2: "known to be complete".

XML parsing: open reading frames

To obtain the nucleotide ORF of a protein sequence from a GenBank XML file, the content of Seq-feat elements, which occur in Bioseq_annot or Bioseq-set_annot elements, have to be interpreted. Seq-feat elements, that contain a Seq-feat_product element, represent a translation product of a DNA sequence. The Seq-interval element contains the nucleotide positions of the ORF that have to be combined to obtain the protein's coding sequence.

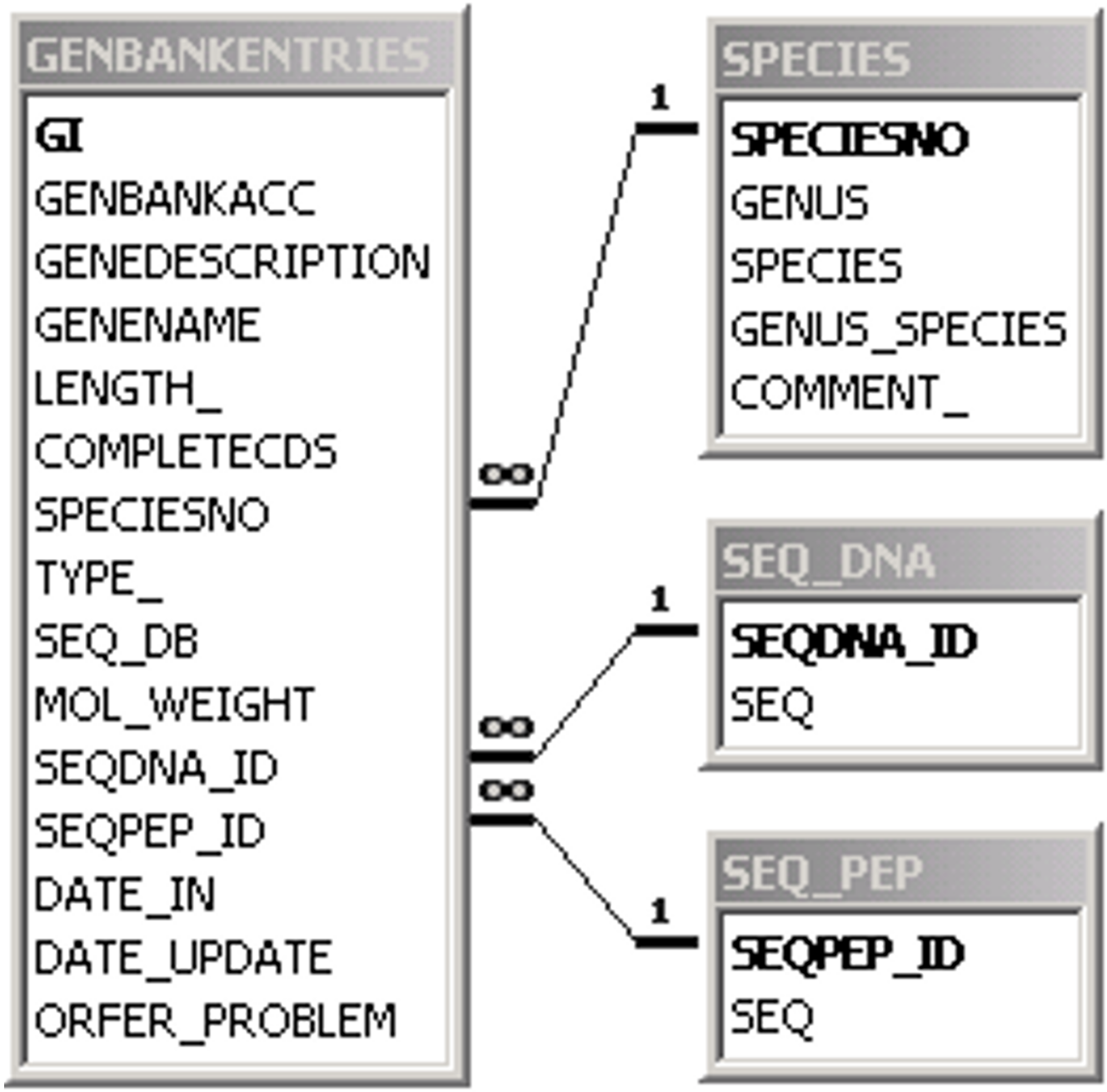


Figure 2
Relational database schema. The database has a main table GENBANKENTRIES, and three associated tables for species information and sequence data.

ORFer only displays and stores ORFs that translate to the protein sequence, starting from the first base pair. If the translation product of the retrieved ORF is not identical to the original protein sequence, two more translations are compared upon introducing a frame-shift. If frame-shifts were introduced to get a match, the sequence entry will be labelled with "ORF was truncated to match protein" and the value 2 will be stored in the field ORFER_PROBLEM in the table GENBANKENTRIES. If none of the three translation is identical to the original protein sequence, the value 1 will be stored in this field. If translation of the sequence fails, e.g. because it contains base pair symbols other than A, C, G, T, the value 3 will be stored. No ORF sequence will be made available by the current version of ORFer, if either translation failed or if the ORF does not translate to the protein sequence.

GenBank contains protein sequences which are annotated to originate from larger precursor proteins. ORFer does not present nucleotide ORFs encoding these protein sequences, but labels them as "Product of larger precursor protein". The value 4 will be stored in the field ORFER_PROBLEM in the table GENBANKENTRIES of a local database.

Non-interactive mode

Since ORFer stores all retrieved sequences in memory, it cannot parse unlimited numbers of sequences per session. For parsing more than about 1,000 sequences, it is recommended to use the non-interactive mode of ORFer, which automatically stores the retrieved sequences in a local database and Fasta text files, but does not keep them in memory.

Retrieval of sets of protein sequences

To test the ORFer application, a set of GI identifiers corresponding to human proteins where retrieved from GenBank using the NCBI Entrez query: "Homo sapiens" [Organism] AND gene_in_genomic [PROP] AND srcdb_genbank [PROP]. The NCBI Entrez query returned 59,274 human protein sequences from the GenBank protein database (June 2002), of which 10,000 were randomly selected for testing. The non-interactive mode of ORFer was used to avoid memory overflow due to the large number of sequences. In the test run, XML files where retrieved for all 10,000 GI identifiers from NCBI. 9,921 sequences were successfully parsed and a protein sequence could be extracted from the XML file. These sequences were entered into a mySQL database. For the remaining 79 sequences, either no protein sequence was found, or the Xerces XML SAX parser reported an error in the structure of the XML file. A tabulator delimited text file of the ORFer output for 1,000 of the 10,000 sequences can be downloaded; see additional file 1.

9,135 of the 9,921 protein sequences were entered into the local database together with their ORF sequence, while 786 sequences were entered without ORF sequence.

No ORF was found in the XML files of 344 of these 786 sequences. These sequences comprise, for example, peptide sequences, for which no ORF is available at GenBank.

For the remaining 442 entries, ORFs either could not be translated by ORFer (180 sequences), or translated to a different sequence than the protein sequence retrieved from GenBank (262 sequences). It was found that the first group of ORFs contained ambiguity nucleotide code, while the latter group of sequences is lacking the last base pair of the last codon. The last amino acid was inferred from the first two base pairs of this codon to generate the GenBank protein sequence.

ORFer uses a BioJava http://www.biojava.org routine for translation of nucleotide sequences. The current version of this routine accepts sequences composed of A, C, G, T only, and will not translate two base pairs at the end of an ORF to an amino acid. If BioJava releases enabling such translations become available, they will be integrated into ORFer.

The proportion of sequence entries tagged as complete and incomplete was determined in the set of 9,921 sequences that ORFer was able to read from GenBank, including protein sequences for which no ORF was determined by ORFer. The proportion of incomplete sequences in this set of protein sequences was found to be quite high, 41.6%. For 38.4%, no information on completeness was found in the annotation. 20.0% of sequences where tagged as complete in GenBank.

Retrieval of protein sequences annotated in genomic sequence entries

ORFer can retrieve complete sets of ORFs and protein sequences annotated in GenBank genomic sequence entries. GenBank sequence entries of cosmid and BAC clones, chromosomes or genomes contain annotation of genes and corresponding ORFs and protein sequences. ORFer can be set to "retrieve all sequences in the XML file". In this mode, when a GI identifier of a genomic sequence is entered, ORFer will present all protein and ORFs found in the respective sequence entry.

A GenBank sequence entry representing a very large sequence, e.g. a complete chromosome, may contain references to smaller genomic sequence entries rather than a long DNA sequence string. ORFer will try to follow these references and retrieve the ORFs and protein sequences included in these smaller sequence entries. Alternatively, the user may download from the NCBI Entrez web site a list of GI identifiers of all the genomic sequences that make up, for example, a chromosome and use this list as input for ORFer.

ORFer was successfully tested with the following genomic sequence entries:

- Bacillus subtilis complete genome, GI 16077068, 4.2 Mbp, 4,112 proteins

- Anopheles gambiae genomic sequence, GI 19612245, 16.4 Mbp, 832 proteins

- Human herpesvirus 5, GI 9625671, 229 kbp, 204 proteins

Conclusions

The ORFer program is an example for parsing the NCBI XML sequence format using the Java programming language. Retrieval of protein sequences with corresponding ORFs was apparently successful for sequences that contain the necessary information. For 91% of human protein sequences in GenBank, ORFer could extract an open reading frame from XML files.

The main benefits of the program are:

• Visual inspection of hundreds of protein or DNA sequence entries in a tabular view.

• Retrieval of ORFs of whole genomes or chromosomes.

• Table view of all proteins annotated in a single DNA sequence entry

• Import of large numbers of sequences from GenBank into a local relational database system

• Export to Fasta sequence files or tab delimited text files

ORFer is a flexible program – it can write data to three different kinds of relational databases, and also to text files in Fasta and tabulator delimited format. It will hopefully prove useful for molecular biologists dealing with larger numbers of DNA or protein sequence.

Materials and Methods

ORFer is freely available from the web site http://www.proteinstrukturfabrik.de/orfer and can be executed via Java Web Start on Windows, Solaris, Linux and Macintosh OS X. ORFer can be run as graphical user interface application or from the command line (Figure 1).

ORFer was entirely written in Java using Borland JBuilder4. The Program uses BioJava libraries http://www.biojava.org for translation of DNA into protein sequences and for writing of Fasta format sequence files. It includes JDBC drivers of Oracle and Sun and the MM MySQL drivers (Mark Matthews). XML parsing is done with Apache Xerces http://xml.apache.org. SQuirrel SQL libraries were used for the copy pop-up menu http://squirrel-sql.sourceforge.net/. A modified version of the TableSorter class of Philip Milne was included for table sorting.

Authors' contributions

SH wrote a first version of the ORFer program and KB wrote the final version. SH and KB conceived of the program. VS and KB designed and implemented the databases.

All authors read and approved the final manuscript.

List of abbreviations

ASN: Abstract Syntax Notation

DBMS: DataBase Management System

JDBC: Java Data Base Connectivity

kbp: Kilo base pairs

Mbp: Mega base pairs

NCBI: National Center for Biotechnology Information

ORF: Open Reading Frame

PCR: Polymerase Chain Reaction

SQL: Structured Query Language

XML: eXtensible Markup Language

Additional material

Additional File 1

ORFer output for 1,000 random human protein sequences as tab-delimited text file.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-3-40-S1.txt]

Acknowledgements

We would like to thank Dr. Grant Langdon for critical reading of the manuscript, Robert Büssow for inspection of the ORFer source code and Prof. Hans Lehrach for his enduring support. The Protein Structure Factory is funded by the German Federal Ministry of Education and Research (BMBF).

References

1. Heinemann U, Frevert J, Hofmann K, Illing G, Maurer C, Oschkinat H and Saenger W **An integrated approach to structural genomics.** *Progress in Biophysics & Molecular Biology* 2000, **73(5):**347-362
2. Michalickova K **Biological sequences and the NCBI toolkit, flatfile I/O and SeqHound API**
3. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I and Clamp M **The Ensembl genome database project.** *Nucleic Acids Research* 2002, **30:**38-41